Ensemble Triangulation for Statistical Machine Translation*

Majid Razmara and **Anoop Sarkar**
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
{razmara,anoop}@sfu.ca

Abstract

State-of-the-art statistical machine translation systems rely heavily on training data and insufficient training data usually results in poor translation quality. One solution to alleviate this problem is *triangulation*. Triangulation uses a third language as a pivot through which another source-target translation system can be built. In this paper, we dynamically create multiple such triangulated systems and combine them using a novel approach called *ensemble decoding*. Experimental results of this approach show significant improvements in the BLEU score over the direct source-target system. Our approach also outperforms a strong linear mixture baseline.

1 Introduction

The objective of current statistical machine translation (SMT) systems is to build cheap and rapid corpus-based SMT systems without involving human translation expertise. Such SMT systems rely heavily on their training data. State-of-the-art SMT systems automatically extract translation rules (e.g. phrase pairs), learn segmentation models, re-ordering models, etc. and find tuning weights solely from data and hence they rely heavily on high quality training data. There are many language pairs for which there is no parallel data or the available data is not sufficiently large to build a reliable SMT system. For example, there is no Chinese-Farsi parallel text, although there exists sufficient parallel data between these two languages and English. For SMT, an important research direction is to improve the quality of translation when there is no, insufficient or poor-quality parallel data between a pair of languages.

*This research was partially supported by an NSERC, Canada (RGPIN: 264905) grant and a Google Faculty Award to the second author.

One approach that has been recently proposed is *triangulation*. Triangulation is the process of translating from a source language to a target language via an intermediate language (aka pivot, or bridge). This is very useful specifically for low-resource languages as SMT systems built using small parallel corpora perform poorly due to data sparsity. In addition, ambiguities in translating from one language into another may disappear if a translation into some other language is available.

One obvious benefit of triangulation is to increase the coverage of the model on the input text. In other words, we can reduce the number of out-of-vocabulary words (OOVs), which are a major cause of poor quality translations, using other paths to the target language. This can be especially helpful when the model is built using a small amount of parallel data.

Figure 1 shows how triangulation can be useful in reducing the number of OOVs when translating from French to English through three pivot languages: Spanish (*es*), German (*de*) and Italian (*it*). The solid lines show the number of OOVs for a direct MT system with regard to a multi-language parallel test set (Section 6.2 contains the details about the data sets) and the dotted lines show the OOVs in the triangulated ($src \rightarrow pvt \rightarrow tgt$) systems. The number of OOVs on triangulated paths can never be less that the first edge (i.e. $src \rightarrow pvt$) and it is usually higher than the second edge (i.e. $pvt \rightarrow tgt$) as well. Thus, the choice of intermediate language is very important in triangulation.

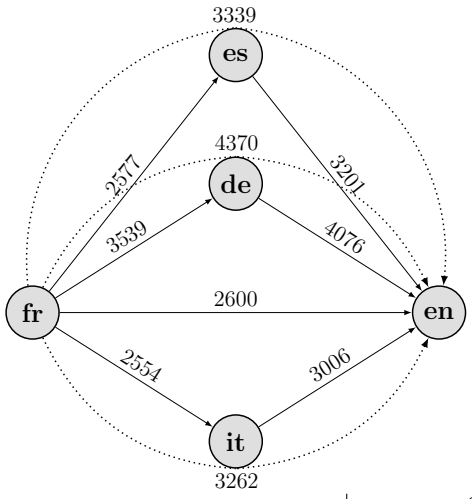
Figure 1 also shows how combining multiple triangulated systems can reduce this number from 2600 (16%) OOVs to 1536 (9%) OOVs. Thus, combining triangulated systems with the original $src \rightarrow tgt$ system is a good idea. When combining multiple systems, the upper bound on the number of OOVs is the minimum among all OOVs in the different triangulations. These OOV rates provide useful hints, among other clues, as to which pivot

International Joint Conference on Natural Language Processing, pages 252–260,
Nagoya, Japan, 14-18 October 2013.

languages will be more useful. In Figure 1, we can expect Italian (*it*) to help more than Spanish (*es*) and both to help more than German (*de*) in translation from French (*fr*) to English (*en*), which we confirmed in our experimental results (Table 1).

In addition to providing translations for otherwise untranslatable phrases, triangulation can find new translations for current phrases. The conditional distributions used for the translation model have been estimated on small amounts of data and hence are not robust due to data sparseness. Using triangulation, these distributions are smoothed and become more reliable as a result.

For each pivot language for which there exists parallel data with the source and the target language, we can create a *src* → *tgt* system by bridging through the pivot language. If there are a number of such pivot languages with corresponding data, we can use mixture approaches to combine them in order to build a stronger model. We propose to apply the ensemble decoding approach of (Razmara et al., 2012) in this triangulation scenario. Ensemble decoding allows us to combine hypotheses from different models dynamically at the decoder. We experimented with 12 different language pairs and 3 pivot languages for each of them. Experimental results of this approach show significant improvements in the BLEU and METEOR scores over the direct source-target system in all the 12 language pairs. We also compare to a strong linear mixture baseline.



direct (*fr-en*)	2600 (16%)
triangulated (*fr*-{*es*, *de*, *it*}-*en*)	2066 (12%)
direct + triangulated	1536 (9%)

Figure 1: Number of OOVs when translating directly from *fr* to *en* (solid lines), triangulating through *es*, *de* or *it* individually (dotted lines), and when combining multiple triangulation systems with the direct system. OOV numbers are based on a multi-language parallel test set and the models are built on small corpora (10k sentence pairs), which are not multi-parallel.

2 Related Work

Use of pivot languages in machine translation dates back to the early days of machine translation. Boitet (1988) discusses the choice of pivot languages, natural or artificial (e.g. interlingua), in machine translation. Schubert (1988) argues that a proper choice for an intermediate language for high-quality machine translation is a natural language due to the inherent lack of expressiveness in artificial languages. Previous work in applying pivot languages in machine translation can be categorized into these divisions:

2.1 System Cascades

In this approach, a *src* → *pvt* translation system translates the source input into the pivot language and a second *pvt* → *tgt* system takes the output of the previous system and translates it into the target language. Utiyama and Isahara (2007) use this approach to triangulate between Spanish, German and French through English. However, instead of using only the best translation, they took the n-best translations and translated them into the target language. MERT (Och, 2003) has been used to tune the weights for the new feature set which consists of *src* → *pvt* and *pvt* → *tgt* feature functions. The highest scoring sentence from the target language is used as the final translation. They showed that using 15 hypotheses in the *pvt* side is generally superior to using only one best hypothesis.

2.2 Corpus Synthesis

Given a *pvt* → *tgt* MT system, one can translate the pivot side of a *src-pvt* parallel corpus into the target language and create a noisy *src-tgt* parallel corpus. This can also be exploited in the other direction, meaning that a *pvt* → *src* MT system can be used to translate the pivot side of a *pvt-tgt* bitext. de Gispert and Marino (2006), for example, translated the Spanish side of an English-Spanish bitext into Catalan using an available Spanish-Catalan SMT system. Then, they built an English-Catalan MT system by training on this new parallel corpus.

2.3 Phrase-Table Triangulation

In this approach, instead of translating the input sentences from a source language to a pivot language and from that to a target language, triangulation is done on the phrase level by triangulating two phrase-tables: *src* → *pvt* and *pvt* → *tgt*:

$$(\bar{f}, \bar{e}) \in T_{\mathcal{F}\to\mathcal{E}} \iff \exists \bar{i} : \ (\bar{f}, \bar{i}) \in T_{\mathcal{F}\to\mathcal{I}} \ \wedge \ (\bar{i}, \bar{e}) \in T_{\mathcal{I}\to\mathcal{E}}$$

where $\bar{f}, \bar{i}$ and $\bar{e}$ are phrases in the source $\mathcal{F}$, pivot $\mathcal{I}$ and target $\mathcal{E}$ languages respectively and T is a set representing a phrase table.

Utiyama and Isahara (2007) also experimented with phrase-table triangulation. They compared both triangulation approaches when using Spanish, French and German as the source and target languages and English as the only pivot language. They showed that phrase-table triangulation is superior to the MT system cascades but both of them did not outperform the direct *src* → *tgt* system.

The phrase-table triangulation approach with multiple pivot languages has been also investigated in several work (Cohn and Lapata, 2007; Wu and Wang, 2007). These triangulated phrase-tables are combined together using linear and log-linear mixture models. They also successfully combined the mixed phrase-table with a *src-tgt* phrase-table to achieve a higher BLEU score.

Bertoldi et al. (2008) formulated phrase triangulation in the decoder where they also consider the phrase-segmentation model between *src-pvt* and the reordering model between *src-tgt*.

Beside machine translation, the use of pivot languages has found applications in other NLP areas. Gollins and Sanderson (2001) used a similar idea in cross-lingual information retrieval where query terms were translated through multiple pivot languages to the target language and the translations are combined to reduce the error. Pivot languages have also been successfully used in inducing translation lexicons (Mann and Yarowsky, 2001) as well as word alignments for resource-poor languages (Kumar et al., 2007; Wang et al., 2006). Callison-Burch et al. (2006) used pivot languages to extract paraphrases for unknown words.

3 Baselines

In this paper, we compare our approach with two baselines. A simple baseline is the direct system between the source and target languages which is trained on the same amount of parallel data as the triangulated ones. In addition, we implemented a phrase-table triangulation method (Cohn and Lapata, 2007; Wu and Wang, 2007; Utiyama and Isahara, 2007). This approach presents a probabilistic formulation for triangulation by marginalizing out the pivot phrases, and factorizing using the chain rule:

$$\begin{aligned} p(\bar{e} \,|\, \bar{f}) &= \sum_{\bar{i}} p(\bar{e}, \bar{i} \,|\, \bar{f}) \\ &= \sum_{\bar{i}} p(\bar{e} \,|\, \bar{i}, \bar{f})\, p(\bar{i} \,|\, \bar{f}) \\ &\approx \sum_{\bar{i}} p(\bar{e} \,|\, \bar{i})\, p(\bar{i} \,|\, \bar{f}) \end{aligned}$$

where $\bar{f}, \bar{e}$ and $\bar{i}$ are phrases in the source, target and intermediate language respectively. In this equation, a conditional independence assumption has been made that source $\bar{f}$ and target phrases $\bar{e}$ are independent given their corresponding pivot phrase(s) $\bar{i}$. The equation requires that all phrases in the *src* → *pvt* direction must also appear in *pvt* → *tgt*. All missing phrases are simply dropped from the final phrase-table.

Using this approach, a triangulated source-target phrase-table is generated for each pivot language. Then, linear and log-linear mixture methods are used to combine these phrase-tables into a single phrase-table in order to be used in the decoder. We implemented the linear mixture approach, since linear mixtures often outperform log-linear ones (Cohn and Lapata, 2007). We then compare the results of these baselines with our approach over multiple language pairs (Section 6.2). In linear mixture models, each feature in the mixture phrase-table is computed as a linear interpolation of corresponding features in the component phrase-tables using a weight vector $\vec{\lambda}$.

$$\begin{aligned} p(\bar{e} \,|\, \bar{f}) &= \sum_i \lambda_i \, p_i(\bar{e} \,|\, \bar{f}) \\ p(\bar{f} \,|\, \bar{e}) &= \sum_i \lambda_i \, p_i(\bar{f} \,|\, \bar{e}) \\ \forall \ \lambda_i > 1 \qquad & \sum_i \lambda_i = 1 \end{aligned}$$

Following Cohn and Lapata (2007), we combined triangulated phrase-tables with uniform weights into a single phrase table and then interpolated it with the phrase-table of the direct system.

4 Ensemble Decoding

SMT log-linear models (Koehn, 2010) find the most likely target language output $\boldsymbol{e}$ given the source language input $\boldsymbol{f}$ using a vector of feature functions $\boldsymbol{\phi}$:

$$p(e|f) \propto \exp\Big(\mathbf{w} \cdot \boldsymbol{\phi}\Big)$$

Ensemble decoding combines several models dynamically at the decoding time. The scores are combined for each partial hypothesis using a user-defined mixture operation $\odot$ over component models.

$$p(e|f) \propto \exp\Big(\mathbf{w_1} \cdot \boldsymbol{\phi_1} \odot \mathbf{w_2} \cdot \boldsymbol{\phi_2} \odot \dots\Big)$$

Razmara et al. (2012) successfully applied ensemble decoding to domain adaptation in SMT and showed that it performed better than approaches that pre-compute linear mixtures of different models. Several mixture operations were proposed, allowing the user to encode belief about the relative strengths of the component models. These mixture operations receive two or more probabilities and return the mixture probability $p(\bar{e} \mid \bar{f})$ for each rule $\bar{f} \rightarrow \bar{e}$ used in the decoder. Different options for these operations are:

- **Weighted Sum (wsum)** is defined as:

$$p(\bar{e} \mid \bar{f}) \;\propto\; \sum_{m}^{M} \lambda_m \exp\left(\mathbf{w}_m \cdot \boldsymbol{\phi}_m\right)$$

  where m denotes the index of component models, M is the total number of them and λ_m is the weight for component m.

- **Weighted Max (wmax)** is defined as:

$$p(\bar{e} \mid \bar{f}) \;\propto\; \max_{m} \left(\lambda_m \exp\left(\mathbf{w}_m \cdot \boldsymbol{\phi}_m\right)\right)$$

- **Model Switching (Switch):** Each cell in the CKY chart is populated only by rules from one of the models and the other models' rules are discarded. Each component model is considered an expert on different spans of the source. A binary indicator function $\delta(\bar{f}, m)$ picks a component model for each span:

$$\delta(\bar{f}, m) = \begin{cases} 1, & m = \underset{n \in M}{\operatorname{argmax}}\ \psi(\bar{f}, n) \\ 0, & \text{otherwise} \end{cases}$$

  The criteria for choosing a model for each cell, $\psi(\bar{f}, n)$, is based on max top score, i.e. for each cell, the model that has the highest weighted best-rule score wins:

$$\psi(\bar{f}, n) = \lambda_n \max_{\bar{e}}(\mathbf{w}_n \cdot \boldsymbol{\phi}_n(\bar{\mathbf{e}}, \bar{\mathbf{f}}))$$

  The probability of each phrase-pair $(\bar{e}, \bar{f})$ is then:

$$p(\bar{e} \mid \bar{f}) = \sum_{m}^{M} \delta(\bar{f}, m)\; p_m(\bar{e} \mid \bar{f})$$

5 Our Approach

5.1 Dynamic Triangulation

Given a *src* → *pvt* and a *pvt* → *tgt* system which are independently trained and tuned on their corresponding parallel data, these two systems can be triangulated dynamically in the decoder.

For each source phrase $\bar{f}$, the decoder consults the *src* → *pvt* system to get its translations on the pivot side $\bar{i}$ with their scores. Consequently, each of these pivot-side translation phrases is queried from the *pvt* → *tgt* system to obtain their translations on the target side with their corresponding scores. Finally a $(\bar{f}, \bar{e})$ pair is constructed from each $(\bar{f}, \bar{i})$ and $(\bar{i}, \bar{e})$ pair, whose score is computed as:

$$p_{\mathcal{I}}(\bar{f} \mid \bar{e}) \;\propto$$

$$\max_{\bar{i}} \exp\Big(\underbrace{w_1 \,.\, \phi_1(\bar{f}, \bar{i})}_{\mathcal{F} \rightarrow \mathcal{I}} \;+\; \underbrace{w_2 \,.\, \phi_2(\bar{i}, \bar{e})}_{\mathcal{I} \rightarrow \mathcal{E}}\Big)$$

This method requires the language model score of the *src* → *pvt* system. However for simplicity we do not use the pivot-side language models and hence the score of the *src* → *pvt* system does not include the language model and word penalty scores. In this formulation for a given source and target phrase pair $(\bar{f}, \bar{e})$, if there are multiple bridging pivot phrases $\bar{i}$, we only use the one that yields the highest score. This is in contrast with previous work where they take the sum over all such pivot phrases (Cohn and Lapata, 2007; Utiyama and Isahara, 2007). We use *max* as it outperforms *sum* in our preliminary experiments.

It is noteworthy that in computing the score for $p_{\mathcal{I}}(\bar{f} \mid \bar{e})$, the scores from *src* → *pvt* and *pvt* → *tgt* are added uniformly. However, there is no reason why this should be the case. Two different weights can be assigned to these two scores to highlight the importance of one against the other one.

A naive implementation of phrase-triangulation in the decoder would require $O(n^2)$ steps for each source sub-span, where n is the average number of translation fan-out (i.e. possible translations) for each phrase. However, since the phrase candidates from both *src* → *pvt* and *pvt* → *tgt* are already sorted, we use a lazy algorithm that reduces the computational complexity to $O(n)$.

5.2 Combining Triangulated Systems

If we can make use of multiple pivot languages, a system can be created on-the-fly for each pivot language by triangulation and these systems can then be combined together in the decoder using *ensemble decoding* discussed in Section 4. Following previous work, these triangulated phrase-tables can also be combined with the direct system to produce a yet stronger model. However, we do not combine them in two steps. Instead, all triangulated systems and the direct one are combined together in a single step.

Ensemble decoding is aware of full model scores when it compares, ranks and prunes hypotheses. This includes the language model, word, phrase and glue rule penalty scores as well as standard phrase-table probabilities.

Since ensemble decoding combines the scores of common hypotheses across multiple systems rather than combining their feature values as in mixture models, it can be used to triangulate heterogeneous systems such as phrase-based, hierarchical phrase-based, and syntax-based with completely different feature types. Considering that ensemble decoding can be used in these diverse scenarios, it offers an attractive alternative to current phrase-table triangulation systems.

5.3 Tuning Component Weights

Component weights control the contribution of each model in the ensemble. A tuning procedure should assign higher weights to the models that produce higher quality translations and lower weights to weak models in order to control their noise propagation in the ensemble. In the ensemble decoder, since we do not have explicit gradient information for the objective function, we use a direct optimizer for tuning. We used *Condor* (Vanden Berghen and Bersini, 2005) which is a publicly available toolkit based on Powell's algorithm.

The ensemble between three triangulated models and a direct one requires tuning in a 4-dimensional space, one for each system. If, on average, the tuner evaluates the decoder n times in each direction in the optimization space, there needs to be n^4 ensemble decoder evaluations, which is very time consuming. Instead, we resorted to a simpler approach for tuning: each triangulated model is separately tuned against the direct model with a fixed weights (we used a weight of 1). In other words, three ensemble models are created, each on a single triangulated model plus the direct one. These ensembles are separately tuned and once completed, these weights comprise the final tuned weights. Thus, the total number of ensemble evaluations reduces from $O(n^4)$ to $O(3n)$.

In addition to this significant complexity reduction, this method enables parallelism in tuning, since the three individual tuning branches can now be run independently. The final tuned weights are not necessarily a local optima and one can run further optimization steps around this point to get to even better solutions which should lead to higher BLEU scores.

6 Experiments & Results

6.1 Experimental Setup

For our experiments, we used the Europarl corpus (v7) (Koehn, 2005) for training sets and ACL/WMT 2005[1] data for dev/test sets (2k sentence pairs) following Cohn and Lapata (2007). Our goal in this paper was to understand how multiple languages can help in triangulation, the improvement in coverage of the unseen data due to triangulation, and the importance of choosing the right languages as pivot languages. Thus, we needed to run experiments on a large number of language pairs, and for each language pair we wanted to work with many pivot languages. To this end, we created small sub-corpora from Europarl by sampling 10,000 sentence pairs and conducted our experiments on them. As we will show, using larger data than this would result in prohibitively large triangulated phrase tables. Table 2 shows the number of words on both sides of used language pairs in our corpora.

The ensemble decoder is built on top of an in-house implementation of a Hiero-style MT system (Chiang, 2005) called Kriya (Sankaran et al., 2012). This Hiero decoder obtains BLEU

[1] http://www.statmt.org/wpt05/mt-shared-task/

src↓		tgt →	en	es	fr
de	pivots	en	–	15.94	13.62
		es	14.47	–	13.43
		fr	14.39	13.45	–
		it	14.14	14.90	11.67
	direct		21.94	20.70	17.37
	mixture		21.86	**22.30**	**18.28**
	wmax		22.49	21.32	18.22
	wsum		22.22	21.42	17.98
	switch		**22.59**	21.80	17.70

src↓		tgt →	de	es	fr
en	pivots	de	–	20.47	17.38
		es	12.95	–	20.78
		fr	14.09	23.25	–
		it	13.00	23.18	19.02
	direct		17.57	28.81	24.58
	mixture		**17.91**	28.89	24.30
	wmax		17.77	29.17	**25.39**
	wsum		17.68	**29.33**	24.70
	switch		17.77	29.32	24.98

src↓		tgt →	de	en	fr
es	pivots	de	–	18.84	23.28
		en	14.50	–	18.55
		fr	12.48	22.81	–
		it	13.69	23.14	23.44
	direct		16.30	28.11	29.83
	mixture		**17.75**	28.99	29.47
	wmax		17.34	**29.23**	**30.54**
	wsum		16.79	28.79	30.12
	switch		16.53	29.16	29.68

src↓		tgt →	de	en	es
fr	pivots	de	–	20.15	22.96
		en	14.84	–	27.84
		es	14.35	23.59	–
		it	14.08	24.08	30.38
	direct		16.56	28.79	35.27
	mixture		17.39	28.83	35.27
	wmax		**17.67**	**29.95**	**36.07**
	wsum		17.41	28.62	35.98
	switch		17.78	28.79	36.33

Table 1: Results of i) single-pivot triangulation; ii) baseline systems including direct systems and linear mixture of triangulated phrase-tables; iii) ensemble triangulation results based on different mixture operations. The mixture and ensemble methods are based on multi-pivot triangulation. These methods are built on 10k sentence-pair corpora.

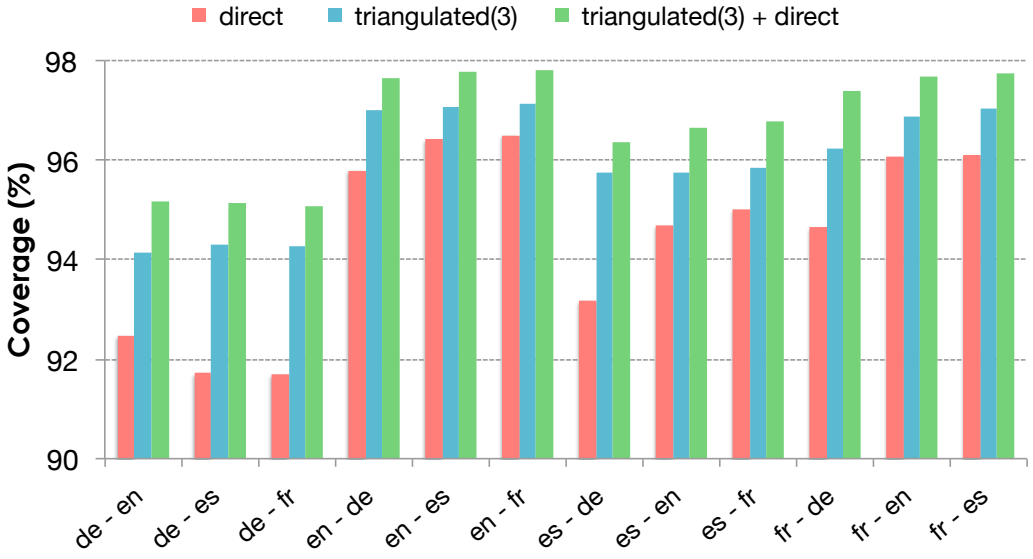
$\mathbf{L}_1$ - $\mathbf{L}_2$	$\mathbf{L}_1$ tokens (K)	$\mathbf{L}_2$ tokens (K)
de - en	232	249
de - es	232	263
de - fr	231	259
de - it	245	253
en - es	250	264
en - fr	251	262
en - it	260	251
es - fr	262	261
es - it	274	252
fr - it	272	251

Table 2: Number of tokens in each language pair in the training data.

scores equal to or better than the state-of-the-art in phrase-based and hierarchical phrase-based translation over a wide variety of language pairs and data sets. It uses the following standard features: forward and backward relative-frequency and lexical TM probabilities; LM; word, phrase and glue-rules penalty. GIZA++ (Och and Ney, 2000) has been used for word alignment with phrase length limit of 10. In both systems, feature weights were optimized using MERT (Och, 2003). We used the target sides of the Europarl corpus ($2M$ sentences) to build 5-gram language models and smooth them using the Kneser-Ney method. We used SRILM (Stolcke, 2002) as the language model toolkit.

6.2 Results

Table 1 shows the BLEU scores when using two languages from $\{fr, en, es, de\}$ as source and target, and the other two languages plus *it* as intermediate languages. The first group of numbers are BLEU scores for triangulated systems through the specified pivot language. For example, translating from *de* to *es* through *en* (i.e. *de* → *en* → *es*) gets 15.94% BLEU score. The second group shows the BLEU scores of the baseline systems including the direct system between the source and target languages and the linear mixture baseline of the three triangulated systems. The BLEU scores of ensemble decoding using different mixture op-



direct	478K	393K	403K	665K	1,084K	1,155K	479K	927K	1,319K	394K	743K	976K
tri + direct	83M	102M	132M	113M	103M	133M	129M	101M	152M	141M	109M	129M

Figure 2: Coverage for i) direct system; ii) combined triangulated system with three 3 languages; and iii) the combination of the triangulated phrase-tables and the direct one. The table shows the number of rules for each system and language pair after filtering based on the source side of the test set.

erations are illustrated at the bottom.

As the table shows, our approach outperforms the direct systems in all the 12 language pairs while the mixture model systems fail to improve over the direct system baseline for some of the language pairs. Our approach also outperforms the mixture models in most cases. Overall, ensemble decoding with *wmax* as mixture operation performs the best among the different systems and baselines. Figure 3 shows the average of the BLEU score of the direct system, mixture models and *wmax* on all 12 systems. On average the *wmax* method obtains 0.33 BLEU points higher than the mixture models.

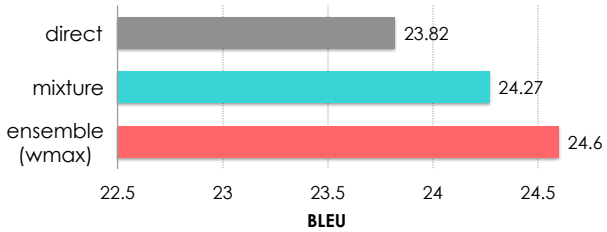


Figure 3: The average BLEU scores of the direct system, mixture models and *wmax* ensemble triangulation approach over all 12 language pairs.

We also computed the Meteor scores (Denkowski and Lavie, 2011) for all systems and the results are summarized in Figure 4. As the figure illustrates, our ensemble decoding approach with *wmax* outperforms the mixture models in 11 of 12 language pairs based on Meteor scores.

6.3 Phrase table coverage

Figure 2 shows the phrase-table coverage of the test set for different language pairs. The coverage is defined as the percentage of unigrams in the source side of the test set for which the corresponding phrase-table has translations for. The first set of bars shows the coverage of the direct systems and the second one shows that of the combined triangulated systems for three pivot languages. Finally, the last set of bars indicate the coverage when the direct phrase-table is combined with the triangulated ones. In all language pairs, the combined triangulated phrase-tables have a higher coverage compared to the direct phrase-tables. As expected, the coverage increases when these two phrase-tables are aggregated. The table below the figure shows the number of rules for each system and language pair after filtering out based on the source side of the test set. This illustrates why running experiments on larger sizes of parallel data is prohibitive for hierarchical phrase-based models.

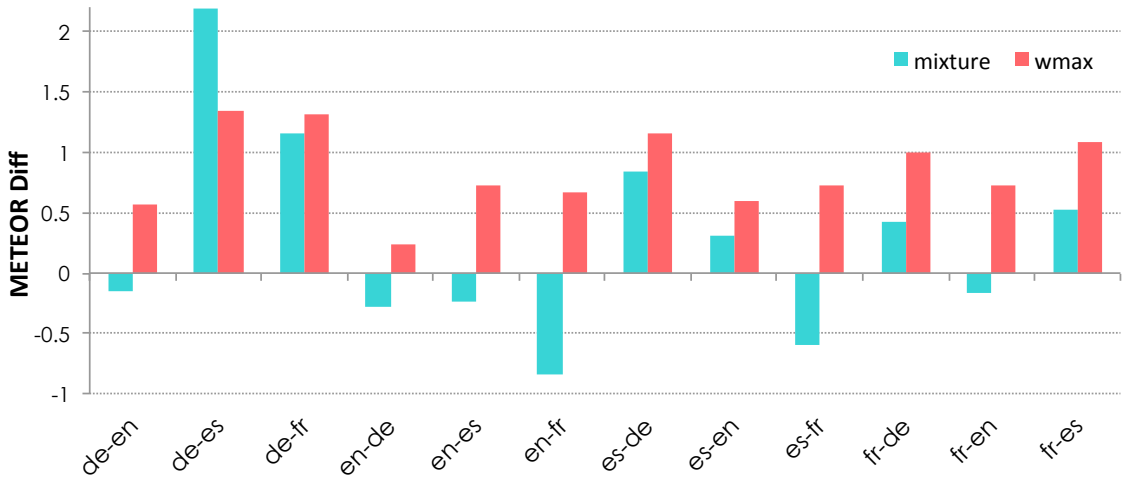


Figure 4: Meteor score difference between mixture models and direct systems as well as the difference between ensemble decoding approach with *wmax* and the direct system.

6.3.1 Choice of Pivot Language

Cohn and Lapata (2007) showed that the pivot language should be close to the source or the target language in order to be effective. For example, when translating between Romance languages (Italian, Spanish, etc.), the pivot language should also be a Romance language. In addition to those findings, based on the results presented in Table 1, here are some observations for these five European languages:

- When translating from or to *de*, *en* is the best pivot language;
- Generally *de* is not a suitable pivot language for any translation pair;
- When translating from *en* to any other language, *fr* is the best pivot;
- *it* is the best intermediate language when translating from *fr* or *es* to other languages; except when translating to *de* for which *en* is the best pivot language (c.f. first finding);

7 Conclusion and Future Work

In the paper, we introduced a novel approach for triangulation which does phrase-table triangulation and model combination on-the-fly in the decoder. Ensemble decoder uses the full hypothesis score for triangulation and combination and hence is able to mix hypotheses from heterogeneous systems.

Another advantage of this method to the phrase-table triangulation approach is that our method is applicable even when there exists no parallel data between source and target languages for tuning because we only use the *src-tgt* tuning set to optimize hyper-parameters, though phrase-table triangulation methods use it to learn MT log-linear feature weights for which having a tuning set is much more essential. Empirical results also showed that this method with *wmax* outperforms the baselines.

Future work includes imposing restrictions on the generated triangulated rules in order to keep only ones that have a strong support from the word alignments. By exploiting such constraints, we can experiment with larger sizes of parallel data. Specifically, a more natural experimental setup for triangulation which we would like to try is to use a small direct system with big *src* → *pvt* and *pvt* → *tgt* systems. This resembles the actual situation for resource-poor language pairs. We will also experiment with higher number of pivot languages.

Currently, most research in this area focuses on triangulation on paths containing only one pivot language. We can also analyze our method when using more languages in the triangulation chain and see whether there would any gain in doing such.

Finally, in current methods all $(\bar{f}, \bar{i})$ phrase pairs of the *src* → *pvt* systems, for which there does not exist any $(\bar{i}, \bar{e})$ pair in *pvt* → *tgt* are simply discarded. However in most cases, such $\bar{i}$ phrases can be segmented into smaller phrases (or rules for Hiero systems) to be triangulated via them. This segmentation is a decoding problem which requires an efficient algorithm to be practical.

References

N. Bertoldi, M. Barbaiani, M. Federico, and R. Cattoni. 2008. Phrase-based statistical machine translation with pivot languages. *Proceeding of IWSLT*, pages 143–149.

C. Boitet. 1988. Pros and cons of the pivot and transfer approaches in multilingual machine translation. *Maxwell et al.(1988)*, pages 93–106.

C. Callison-Burch, P. Koehn, and M. Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 17–24. Association for Computational Linguistics.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270, Morristown, NJ, USA. ACL.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, Prague, Czech Republic, June. Association for Computational Linguistics.

A. de Gispert and J.B. Marino. 2006. Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In *Proc. of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 65–68.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

T. Gollins and M. Sanderson. 2001. Improving cross language retrieval with triangulated translation. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 90–95. ACM.

Martin Kay. 1997. The proper place of men and machines in language translation. *Machine Translation*, 12(1/2):3–23.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Shankar Kumar, Franz Josef Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. In *EMNLP-CoNLL*, pages 42–50. ACL.

Gideon S. Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8, Stroudsburg, PA, USA.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the ACL*, pages 440–447, Hongkong, China, October.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the 41th Annual Meeting of the ACL*, Sapporo, July. ACL.

Majid Razmara, George Foster, Baskaran Sankaran, and Anoop Sarkar. 2012. Mixing multiple translation models in statistical machine translation. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*, pages 940–949. The Association for Computer Linguistics.

Baskaran Sankaran, Majid Razmara, and Anoop Sarkar. 2012. Kriya – an end-to-end hierarchical phrase-based mt system. *The Prague Bulletin of Mathematical Linguistics*, 97(97), April.

K. Schubert. 1988. Implicitness as a guiding principle in machine translation. In *Proceedings of the 12th conference on Computational linguistics-Volume 2*, pages 599–601. Association for Computational Linguistics.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286.

M. Utiyama and H. Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proceedings of NAACL-HLT*, volume 7, pages 484–491.

Frank Vanden Berghen and Hugues Bersini. 2005. CONDOR, a new parallel, constrained extension of powell's UOBYQA algorithm: Experimental results and comparison with the DFO algorithm. *Journal of Computational and Applied Mathematics*, 181:157–175, September.

H. Wang, H. Wu, and Z. Liu. 2006. Word alignment for languages with scarce resources using bilingual corpora of other language pairs. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 874–881. Association for Computational Linguistics.

H. Wu and H. Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.